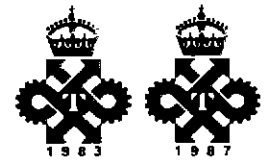
BBC RD 1987/5



BBC

Research Department Report

July 1987

A REVIEW OF THE SEMICONDUCTOR STORAGE OF TELEVISION SIGNALS:

Part 1 — Historical introduction and design philosophy

J.L. Riley, M.Sc.(Eng), C.Eng., M.I.E.E.

Research Department, Engineering Division
THE BRITISH BROADCASTING CORPORATION

BBC RD 1987/5

UDC 621.377.62.25 : 621.397

A REVIEW OF THE SEMICONDUCTOR STORAGE OF TELEVISION SIGNALS :

Part 1 — Historical Introduction and Design Philosophy

J.L. Riley, M.Sc.(Eng), C.Eng., M.I.E.E.

Summary

The emerging semiconductor memory technology over the last two decades has seen an accelerated growth in memory chip density and capacity against a background of falling costs in terms of pence per bit. In this, the first of two Reports, the trends of this technology and some of the important operational characteristics of each ensuing generation of device are described. The design philosophy for forming the devices into useful tools for the storage of television signals is also outlined. In the second and companion Report, some of the applications, as developed in BBC Research Department over the period 1975 - 1986, are described in detail. These include improved television synchronisers, high-quality PAL decoders, television noise reducers, film-dirt concealment equipment and buffer storage for television picture-processing equipment such as stills stores.

Issued under the Authority of

M.E.B. Moffat

Head of Research Department

**Research Department, Engineering Division,
BRITISH BROADCASTING CORPORATION**

(EL-185)

July 1987

It uses SI units in accordance with B.S. document
PD 5686

A REVIEW OF THE SEMICONDUCTOR STORAGE OF TELEVISION SIGNALS:

Part 1 — Historical Introduction and Design Philosophy

J.L. Riley, M.Sc.(Eng), C.Eng., M.I.E.E.

A REVIEW OF THE SEMICONDUCTOR STORAGE OF TELEVISION SIGNALS:

Part 1 — Historical Introduction and Design Philosophy

J.L. Riley, M.Sc.(Eng), C.Eng., M.I.E.E.

1. INTRODUCTION

The falling cost, increased density and capacity, and widespread availability of semiconductor memory devices makes the use of television picture stores a viable proposition for a wide range of applications. These include the construction of delay elements for digital filters used in standards conversion, PAL decoding, television synchronisers and noise reducers. They also include the storage of television pictures as randomly-accessible 'stills' in a mass store. In such cases the number of data samples required to define a picture is in the region of half-a-million.

Other applications require considerably less storage. For example, for filters employing television line delays or for the storage of teletext pages, only a few thousand samples need be stored. Smaller capacity devices are available which have the advantage, over their bigger brothers, of speed of access and ease of use as will be described later.

This Report, the first of a pair, discusses the emerging technology of semiconductor memory and describes some of its salient features. Some design factors to be taken into account when using such devices in practical stores are outlined. The second Report[1] describes a range of applications to digital television engineering developed at BBC Research Department over the past ten years.

2. THE EMERGING SEMICONDUCTOR MEMORY TECHNOLOGY

2.1 Early devices

The Computer Industry, over the last two decades at least, has grown around and derived its momentum from a rapidly developing memory technology. At first there were vacuum tubes and then ferrite cores in the 1950s and 1960s. Semiconductor memory devices followed and continue to the present day. So far, over this period there has been a tremendous rate of development which has seen the chips grow from tiny 64-bit devices to a massive 1 Mbit capacity, from devices which were difficult to use to ones which are comparatively easy.

The whole range of memory devices now available is categorised by cost and speed performance in the graph shown in Fig. 1. Semiconductor memory encompasses on the one hand the fastest, and yet costliest, devices (ECL, I^2L and TTL) and on the other the medium speed, medium cost (dynamic MOS) devices with sub-100 ns access times giving about 1000 bits a penny. Slower but larger devices based on charge-coupled device (CCD) technology were for a time considered to be the future high-density, low-cost memory medium but early attempts to build reliable devices foundered. Bubble memory promised to fill this need but access times are presently very slow. At the cheapest, but certainly the slowest, end of the range lie the moveable-head magnetic devices although these are now facing stiff competition from semiconductor memory as the technology develops.

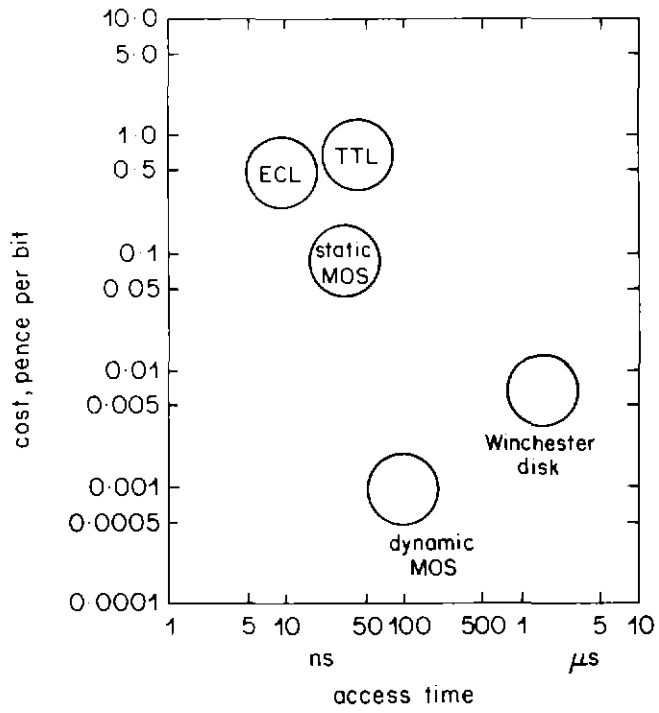


Fig. 1 - Cost-performance characteristic of several types of memory device

Applications within the computer industry diversified in the 1980s. The widespread use of microprocessors demands lower power consumption whilst their use with scanned displays demands faster access. This has led to a continued impetus for semiconductor memory development and in particular the current emergence of CMOS technology.

The chief concern in this Report is with metal-oxide-silicon (MOS) semiconductor memory which can be broadly divided into two groups and referred to as 'static' and 'dynamic' because of the inherent cell structure. In this Section, the important milestones in the development of MOS technology are described in order to explain the trends in memory density, access time, power consumption and cost. Attention is drawn to the problems encountered in constructing such devices and how they were overcome. Each new generation has shown some improvement over its predecessor which makes for faster access, lower power-consumption and lower cost per bit.

Some of the first semiconductor memory devices to emerge were the 64-bit and 256-bit ones from Intel which contained memory cells arranged in an orthogonal two-dimensional matrix as shown in Fig. 2. Each cell was identified by a particular row and column address number and contained a digital binary digit (bit) of information which was either a '1' or a '0'. The individual cells of these devices were based on a cross-coupled flip-flop comprising, in most cases, six MOS transistors as shown in Fig. 3. Data is stored twice, in its true and complement forms, and this gives rise to pairs of column (bit) lines. Each cell is accessed by applying a voltage to a selected row (word) line so that the appropriate switch transistors, A and D, are turned on to connect one cell to each column. One column is selected by applying a voltage to the column switch transistors, G and H, to connect a single cell to the write and read transistor switches, I and J. In this particular design the data is written in true form and read from the complemented data line and transistors B and C form the loads for the flip-flop. This type of cell forms the basis of all existing static memory devices.

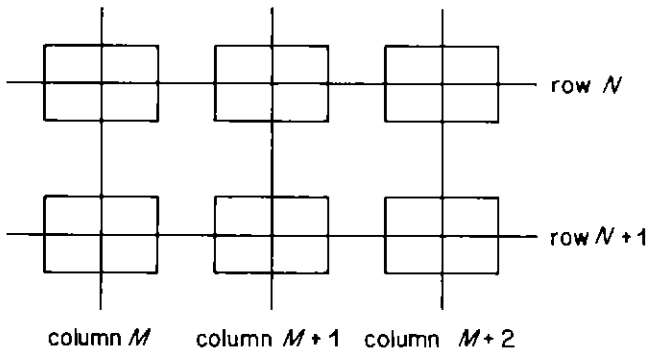


Fig. 2 - Memory cell matrix, showing memory cells addressed by a ROW and COLUMN address count.

These early devices were simple in structure but they were of comparatively large size and had a high power consumption. There was no decoding of the addresses on the chip. The MOS transistors were built using P-channel technology (P-MOS) which was more easily controlled in the early days. The resulting high cost promoted few serious applications but by the end of the 1960's a new generation 1024-bit device became available from Intel.

To overcome the large size and power-consumption disadvantages of the 256-bit devices, the 'dynamic' memory cell was devised. The static cell was replaced by a capacitor holding a stored charge and transistor switches to connect it to the memory matrix. In the 1 Kbit Intel 1103 device each memory cell consisted of three P-MOS transistors, the gate-to-source capacitance of one transistor forming the cell capacitor. The problem with this arrangement is that the stored charge will eventually leak away through the finite resistance associated with the gate of this transistor to ground. This would typically occur within 2 ms. Special arrangements have to be made to sense the charge on each cell capacitor within this time and 'refresh' it to the original condition. This is normally performed row-by-row so that to refresh the entire matrix array, a refresh operation is applied to each

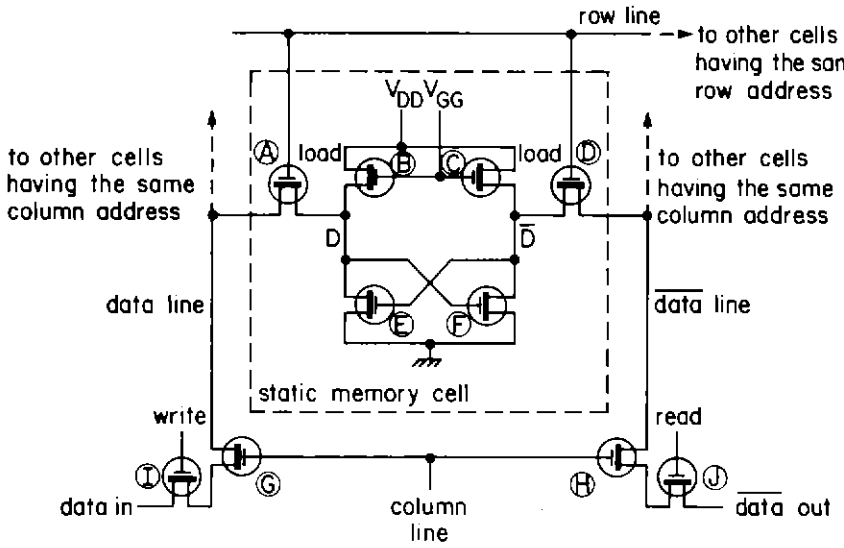


Fig. 3 - Basic static read-write memory cell.

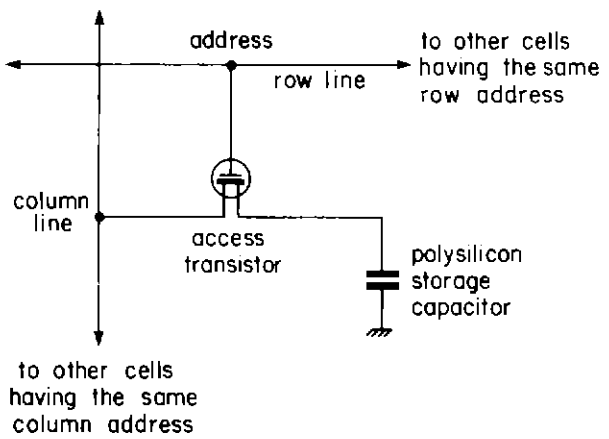
row within 2 ms. The benefits of smaller cell size and lower power consumption more than outweighed the necessity to refresh.

The device was not without its drawbacks, however, and another 1 K bit dynamic memory (DRAM), the MK4006 from Mostek, improved upon it and allowed the user to supply TTL level clocks with minimal timing considerations. Also, for the first time, the address decoding circuits were incorporated within the chip. Memory access times were in the order of 350 ns from the supplied address clock.

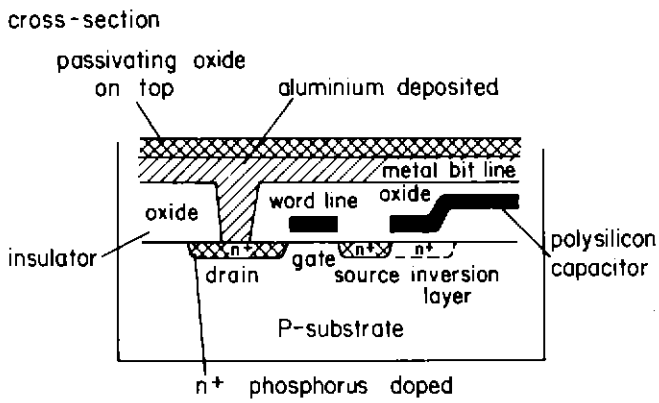
2.2 The 4 Kbit generation

The introduction of the next generation, 4 K x 1, dynamic memory chips produced a variety of designs which competed with one another. There were at least five major designs comprising two differently-arranged 22-pin dual-in-line packaged (DIL) devices, two differently-arranged 18-pin DIL devices and a revolutionary 16-pin DIL device! Access times, power dissipation and chip size varied over a range of more than 2:1. The smaller 16-pin design was achieved by time-multiplexing the 12-bit address onto 6 signal pins so that the row-address and column-address components had to be supplied separately with independent clocks, referred to as the row address strobe (RAS) and column address strobe (CAS) respectively. The advantage of greater packing density compensated for the more awkward addressing arrangements and slightly inferior access times. On the later MK 4027 devices, internal circuitry allowed the complex timing of the row and column addresses to be handled on the chip making it relatively easy for the user to drive. Access times came down to about 150 ns.

The memory cell was eventually reduced to a capacitor and a single transistor as shown in Fig. 4(a) and N-MOS technology was adopted as techniques developed. N-MOS is better because the threshold voltages are lower — suiting TTL compatibility —



(a)



(b)

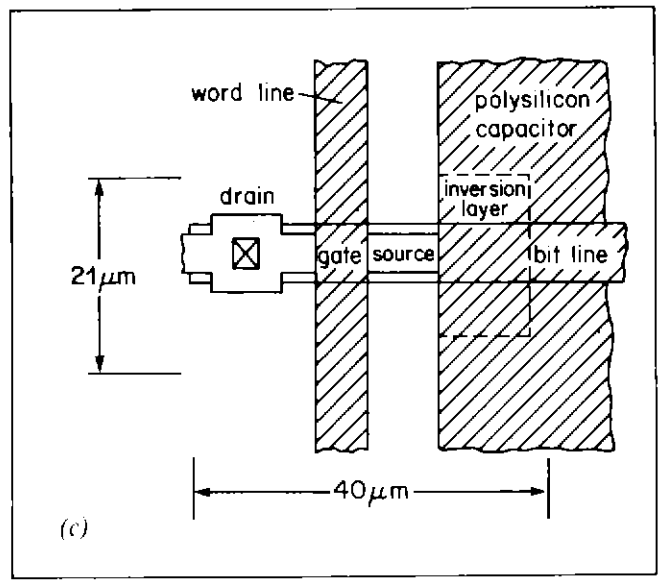


Fig. 4 - The single polysilicon process for 4 K x 1 devices: (a) circuit (b) cross-section (c) plan view

and the greater mobility of the electron carriers provides an inherently faster device. A low-resistance polysilicon material (POLY) for the gates and row lines replaced the aluminium used previously. The cross-section view shown in Fig. 4(b) is accompanied by a plan view of the POLY I process in Fig. 4(c) to illustrate the compact cell arrangement. The smaller cell size and hence capacitor size (about 0.07 pF) meant that the voltages required to be sensed in the refresh operation were also correspondingly lower. These voltages are an attenuated version of the signal from the cell because of the capacitative divider action of the cell capacitance and the stray capacitance of the column lines — the latter may be measured in picofarads. One of the technological hurdles overcome, was to develop a sense amplifier design which could cope with these low voltage levels. Other problems remaining included the high power dissipation — largely due to the extra circuitry required to support the memory — inadequate noise margins and unexplained (at that time) 'soft' errors which are those

produced randomly from other than physical defects on the chip but which can be recovered by re-programming the data.

2.3 The 16 Kbit generation

The next generation DRAM, which appeared in 1976 (the 16 K x 1), attempted to overcome the drawbacks of the 4 K x 1 device. The new technology, which permitted a greater memory density, incorporated two separate layers of polysilicon (POLY II) in the cell construction (shown in Fig. 5) and located the capacitor below the transistor. This technology was originally developed for CCD chips with minimum line widths down to 5 μm and resulted in cell sizes of less than 500 μm^2.

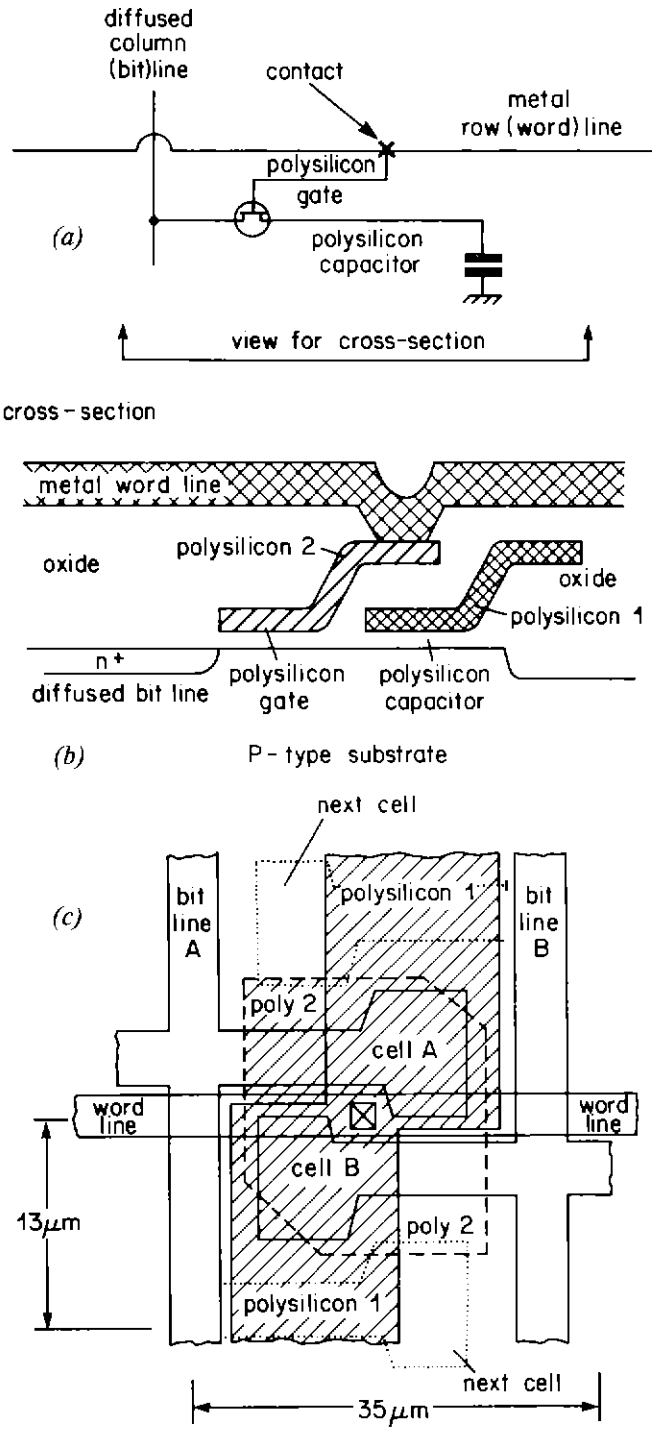


Fig. 5 - The double polysilicon process: (a) circuit (b) cross-section (c) plan view

At this time there was considerable standardisation in the device packaging and performance so that to a large extent devices were interchangeable and effectively multi-sourced — a fact which played an important part in assisting costs to fall. The 16-pin DIL package became universal, the extra seventh address pin being found by taking over a pin previously used for a chip select function. TTL compatibility was also important and the problems of high-power dissipation and inadequate noise margins were largely overcome by adopting balanced amplifier designs. The output driving specification was increased to permit two TTL loads and up to 100 pF to be handled[2]. The soft errors were found to be caused by alpha-particles emanating from the chip carrier casing and bombarding the chip — the result being to cause spurious electron-hole pairs to be generated and upset the stored charge[3]. Special coatings over the chip helped to reduce this effect to an acceptable level.

Once the single transistor cell had arrived, further advances in the technology relied on reducing the cell size. There are several ways to effect this reduction without altering the basic design. Throughout the 4 K and 16 K development period cell sizes were progressively reduced in two dimensions in what was termed a 'shrinking' operation. At the transistor level, the length-to-width ratio of the channel determines its characteristics including resistive properties, gain, speed performance and relative size. Photographic size reduction produces a smaller device which has similar properties to the original provided the length-to-width ratio is kept constant.

A size reduction involving all three dimensions is called 'scaling' and it is this technique which is responsible for the move to 64 K DRAMs and the later generation 16 K devices. Referring to Fig. 6 and Table 1, an example is given of a transistor fabricated in the POLY II technology and scaled by a factor K.

Since the field strength is required to be kept constant the voltage also scales by K: the device area is reduced by a factor of K^2 and the stored charge by a factor K. Both these reduce the transit time and increase the speed performance. The cell current is reduced by a factor K and hence power dissipation by a factor K^2. As both power and voltage are lower the reliability is improved. Scaling techniques are limited by the tolerances of the photo-lithographic equipment employed to make the masks.

Table 1
Example of Scaled Technology

	Scale factor	POLY II	Scaled POLY
Channel length μm	K	5	2.1
Power supply voltage	K	12	5
Junction depth μm	K	1.2	0.5
Oxide depth Å	K	850	354
Cell area μm^2	K^2	600	105
Capacitance pF	K	0.07	0.03
Power dissipation μW	K^2	40	7

$$K = \frac{5}{12}$$

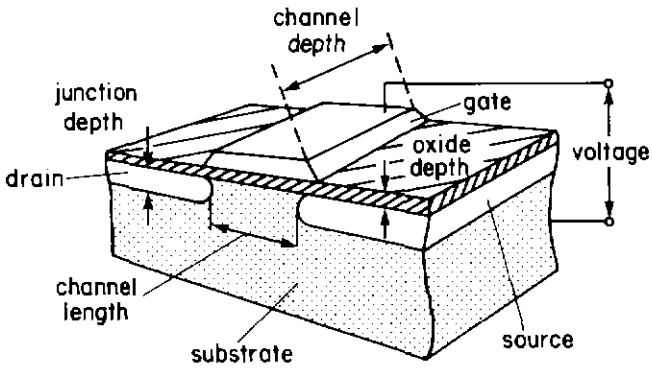


Fig. 6 - Simplified three-dimensional diagram of MOS transistor.

2.4 The 64 Kbit generation

The 64 K DRAM was achieved by the results of several technological developments coming together[4]. The scaled N-MOS mechanism made TTL compatibility easier to achieve and it allowed the +12 V supply to be dispensed with. Substrate bias circuits generated the negative voltage required by the substrate by using the output of a ring oscillator capacitively coupled to the substrate. It then became practical to use a single 5 V power supply instead of three (±5 V, +12 V). This released two of the three package pins for further address lines and opened the way to extend the now standard 16-pin DIL package up to 256 K DRAM devices. In many cases, pin 1 on 64 K devices was left unconnected to anticipate this — see Fig. 7. The relative sizes of an INMOS 64 K DRAM chip and its DIL package are illustrated in Fig. 8.

As the device dimensions are reduced the stored charge is correspondingly reduced and this has two effects. The signal voltage available to the sense amplifiers is decreased making the sense amplifier design more critical. A method of folding the column lines back on themselves helped to reduce the stray capacitance and reduce the charge attenuation factor. Secondly, the difference between the number of electrons sensed as a '1' and those sensed as a '0' is smaller. Since the number of electron-hole pairs produced within the silicon by an incident alpha-particle remains substantially constant, the probability of an alpha-particle error is increased. Package materials were therefore improved and protective top-coats increased to reduce this problem. Fabricating the column lines in metal rather than by diffusion reduced the exposed silicon area and also helped to increase alpha-immunity.

Another problem arises when the channel length is less than about 3 μm: the source and drain regions are so close together that their respective depletion regions within the silicon may overlap causing an unwanted current path between source and drain which is out of control of the gate. This problem was tackled by a method of ion implantation whereby the diffusions of source and drain are graded to minimise the effects of the unwanted path.

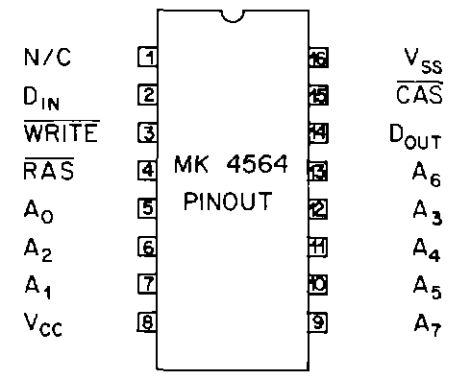


Fig. 7 - Pin layout for MK4564 64 K dynamic memory chip.

2.5 The 256 Kbit generation and beyond

The scaled N-MOS technology continued through several levels of scaling until line widths were down to about 2 μm and this was used for the first 256 K DRAMs. At this density, one of the outstanding difficulties is achieving a satisfactory yield in the manufacturing process. To this end a number of spare columns of cells and row decoding circuitry are normally integrated into the chip and subsequently selected and re-addressed to replace defective ones prior to the final interconnection.

Recently there has been a move away from N-MOS to an advanced complementary metal oxide silicon (CHMOS) technology with its inherently lower power consumption resulting from balanced circuit design where no steady direct current is drawn[5, 6, 7, 8]. CHMOS technology combines all the advances made by scaled N-MOS technology with CMOS circuit

design. The P-MOS transistors and memory cells are contained within n+ wells within the P-type substrate — see Fig. 9: this 'burying' of the memory cell within the silicon improves the alpha-immunity. The smaller scale achieved (1.2 μm line widths) increases the device speed so that sub-100 ns access times are common. It is claimed that CHMOS parts can match N-MOS ones for speed performance and can directly replace them with power savings.

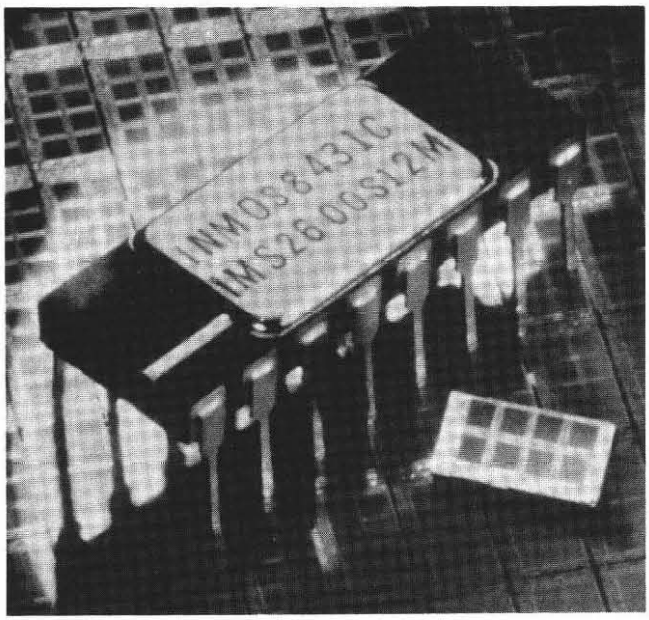


Fig. 8 - An INMOS 64 K DRAM showing the relative sizes of DIL package and chip area

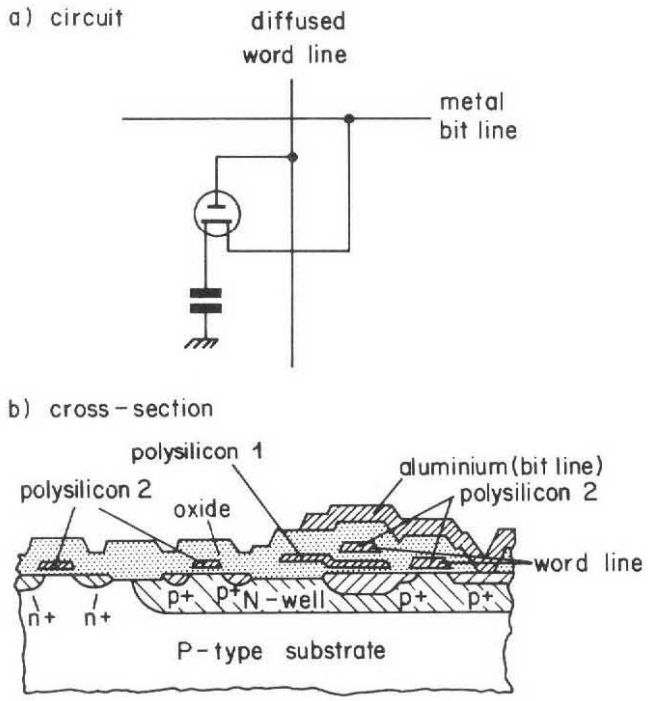


Fig. 9 - The advanced CMOS process: (a) circuit (b) cross-section

Several manufacturers are now involved in 1 Mbit DRAM development with volume production expected well before the 1990s. The technology depends on deeper entrenched cell capacitors where the surface area of the capacitor is effectively increased by utilising the side-walls of the trench as well as its base. This approach also helps to improve alpha-immunity. The problems yet to be overcome include the ability to achieve trenches of reasonable depth and profile and to make oxide layers thin enough to match the scaling. Looking even further ahead, 4 Mbit and 16 Mbit devices are expected to require trench depths down to 7 μm deep with 0.5 μm design rules.

2.6 Developing trends of the technology

From the last Section it is clear that there are several trends in semiconductor memory development which are worth summarising. These are essentially based on historical evidence and can be projected into the future with some degree of confidence. The important technological features of minimum line width, cell area, chip size, access time and power consumption are considered individually. A performance trend can be well illustrated in the speed-power product. Finally the costs are discussed in terms of per unit cell. The statistics presented relate to dynamic memory primarily although some parameters relate also to static devices.

One distinct trend which has characterised the evolution of dynamic memory chips is the quadrupling of storage capacity which has continued at the rate of a new generation about every four years since the 1970s and looks set to continue at an even greater rate for the rest of the century. The rapid build-up of the quantity of devices shipped has been fuelled by the success of the previous generations and stiff competition between manufacturers. The total number of bits taken worldwide is very nearly equivalent to a doubling every year.

The most advanced technologies tend to appear first in dynamic memory devices — see Fig. 10 — although there have been exceptions such as the CCD development which spawned the POLY II process and several programmable read-only memory devices (PROMs) which some manufacturers treated as trial devices for the scaled POLY technology.

Each generation of device is often produced by more than one technology: some manufacturers are content to use the existing technology pushed to its limits to produce an early device ahead of the competition while others are keen to develop the new technology and beat the competition on performance later. In this way the technology is constantly under review and as each device is introduced, another is on the way.

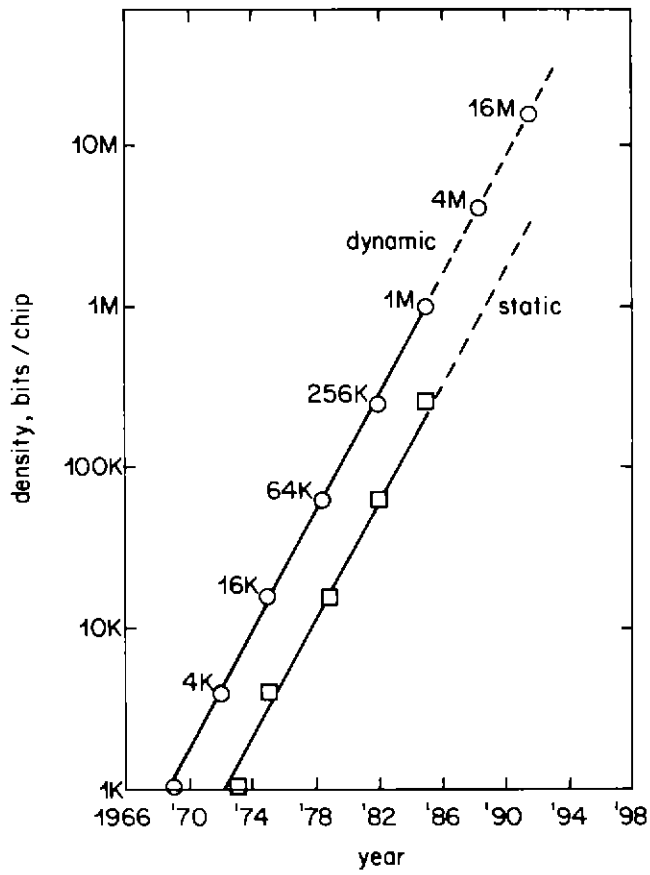


Fig. 10 - Semiconductor memory development

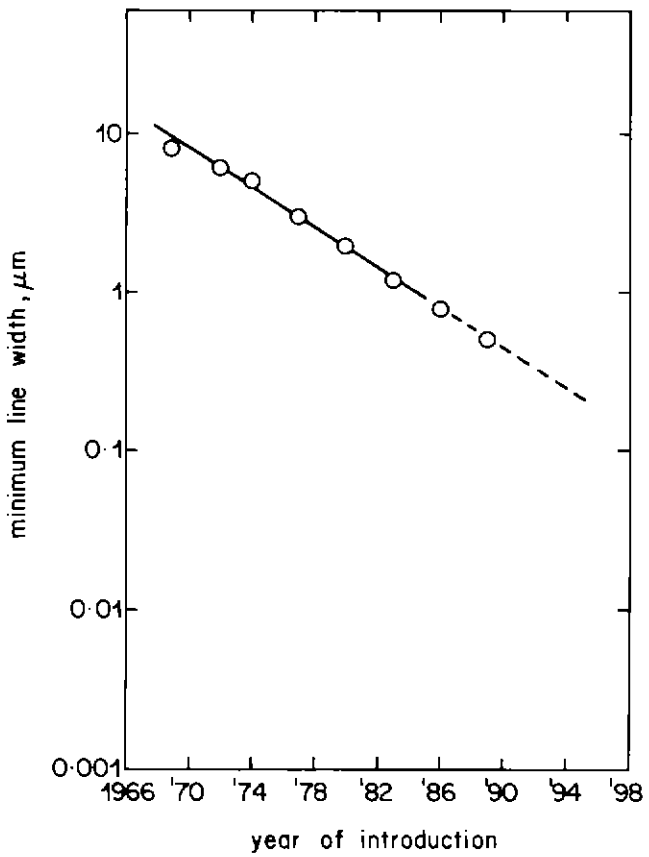


Fig. 11 - The trend in dynamic memory minimum line width

The minimum line width — see Fig. 11 — has already been reduced by a factor of 10 since the early memory devices appeared. With line widths approaching a micron the limits of conventional photolithographic processes are reached and the future lies in electron beam and, later, X-ray techniques. Associated with line width is the overall cell area — see Fig. 12. The 1 Mbit chips are expected to achieve a cell area of 32 μm^2. The overall chip or die size has varied, generally between about 30 mm^2 down to 10 mm^2 depending on the memory capacity and technology used. The chip size affects system costs because the smaller the chip, the cheaper it is to produce and sell and also the device yield is improved.

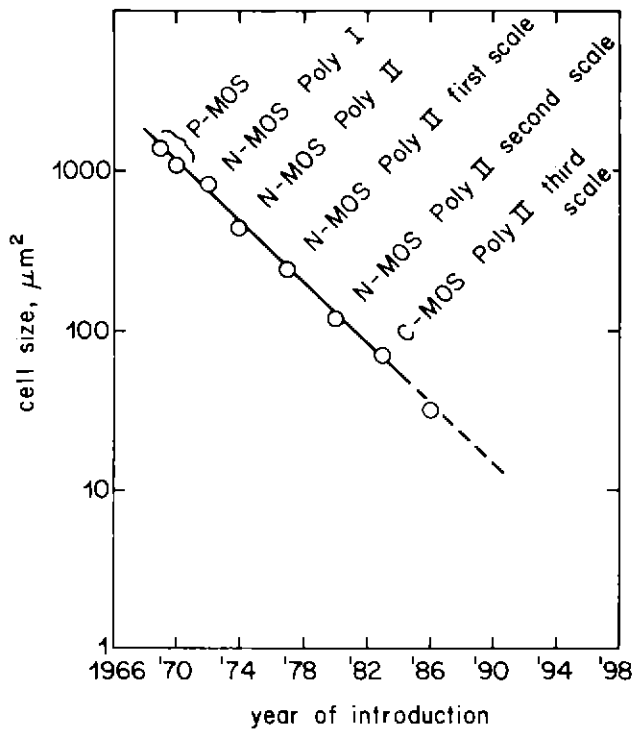


Fig. 12 - The trend in dynamic memory cell size.

The memory access time has decreased but not as rapidly as other factors. The multiplexed addressing arrangement has always penalised the access times of dynamic memories but smaller scaled cells and lower-resistance column-lines and row-lines have helped to reduce on-chip signal delays. With sub-micron geometries and now CMOS circuitry, power consumption levels are set to fall below 1 μW/bit for the first time, making the 256 K DRAM chip run several times cooler than a 1 K DRAM chip fifteen years ago. A useful overall figure-of-merit, combining a measure of speed and power performance is the power-delay product or power/speed ratio shown in Fig. 13, measured in pico-joules.

The trend of dynamic memory development is such that as the package capacity increases the initial cost per bit is more than its immediate predecessor but

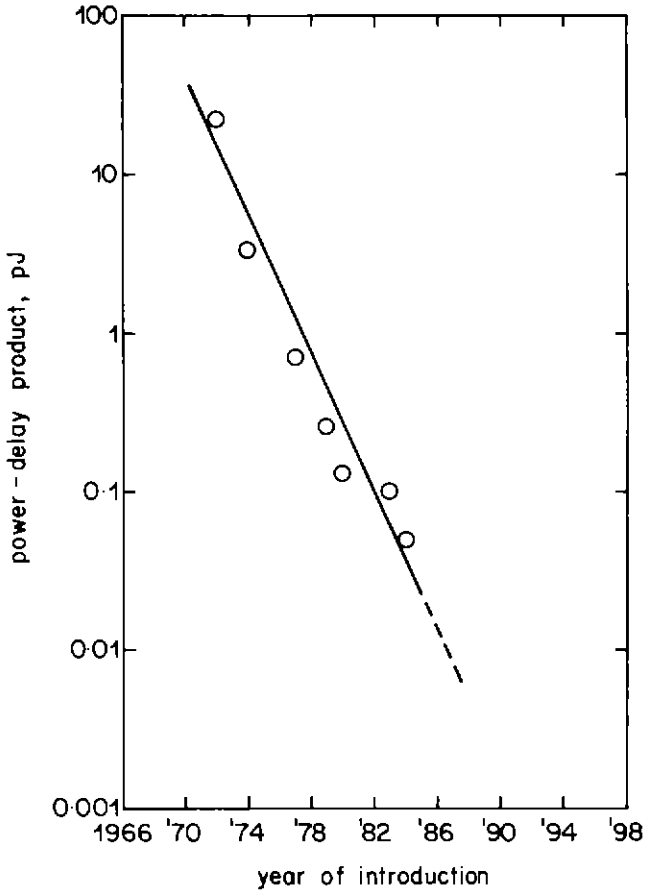


Fig. 13 - The trend in dynamic memory power-delay product

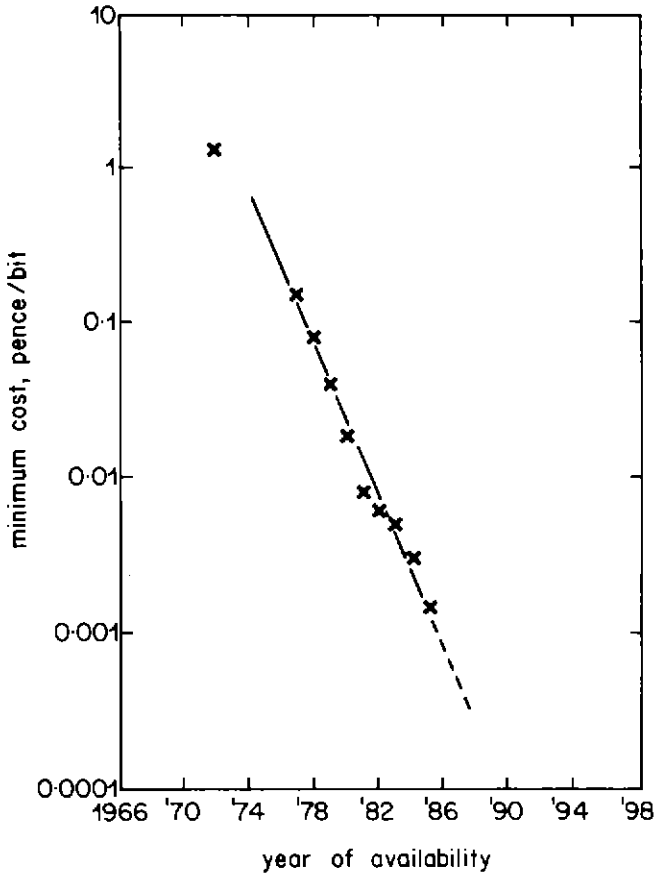


Fig. 14 - The trend in dynamic memory costs.

will, at some time in its life, become competitive and eventually 'cross-over' and therefore cost less per bit until it is succeeded by the following generation. The cost of implementing a digital system based on these devices becomes competitive sooner than might be expected as the higher density allows a reduced package count and reduced power-supply requirements. The figures presented in Fig. 14 ignore the trends of individual devices and concentrate on the best value which was available at any time. These historic costs are based on the market rates pertaining to the UK in quantities of more than 1000 units. The steady fall in unit costs has continued by about a factor of ten every five years.

3. FEATURES OF DYNAMIC SEMICONDUCTOR MEMORY

3.1 Multiplexed addressing and basic read cycle

Traditionally, the memory address has been multiplexed into its row and column components in order to contain the pin count of the memory chip package. Each component is latched into the memory device by independent address strobes and the precise timing of these is critical if the memory access time is to be minimised. The chip designers have attempted to simplify this timing problem by defining the timing relationships, as far as possible, on the chip itself. To help understand this, a simplified functional block diagram of a typical dynamic read-write memory device is presented in Fig. 15.

The figure shows the applied memory address as an n-bit wide signal which is time-multiplexed to be alternately the row component and the column component of the address. For example, a 16 K device, requiring 14 data bits to define it, has an n value of seven. The first action required to access the device is to apply the row address — see Fig. 16(a) — and as soon as the row address inputs are valid (after a period t_{ASR}) the first address strobe may be activated. This strobe is referred to as the row address strobe (RAS) and is an active low signal. It is responsible for initiating a variety of memory cycles which, once begun, must not be aborted.

The falling edge of $\overline{\text{RAS}}$ triggers an internally generated clock which performs three further functions. The first of these is to latch the row address into the chip and decode it. Secondly, the selected row is enabled and data is destructively read from each cell in the selected row by dumping its charge onto its respective column sense line. A sense amplifier for each column detects the change in voltage level on the column line as a result of the deposited charge and the signal is amplified. The third function is to latch the

data into these sense amplifiers. The amplified signals are fed back onto the column sense lines, thus restoring (refreshing) the cells to their original voltages. At this time the sense amplifiers contain the same data as in the selected row — and this remains so until $\overline{RAS}$ is de-activated. The minimum active period for $\overline{RAS}$ is necessary to allow the sense amplifiers time to restore the data, (t_{RAS}).

has become active. If $\overline{CAS}$ is applied beyond the t_{RCD} (max) limit the access time is exclusively determined by $\overline{CAS}$ (t_{CAC}) rather than $\overline{RAS}$ (t_{RAC}). The output buffer is enabled by the $\overline{CAS}$ generated clock and this effectively completes the read access of the memory device. This buffer remains enabled until $\overline{CAS}$ becomes inactive. Before another access can occur $\overline{RAS}$ must be held inactive for a prescribed precharge

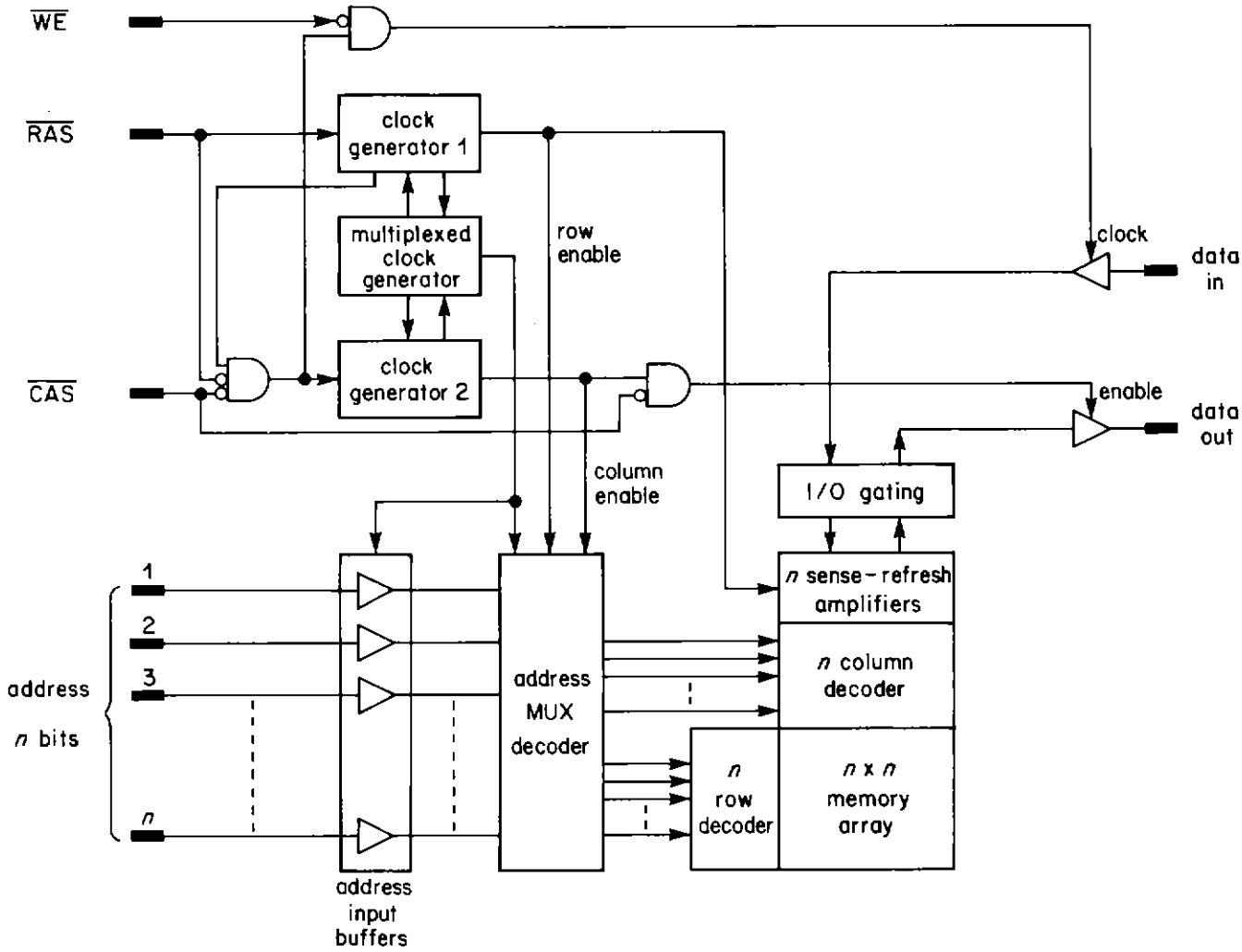


Fig. 15 - A simplified functional block diagram of a dynamic read-write memory device (signals only shown)

Once the row address hold time (t_{RAH}) has been met the column address may be applied and as soon as this is valid (after a period t_{ASC}) the second address strobe may be activated. As soon as this column address strobe (CAS) is applied the data output buffer is immediately disabled and the data output assumes a high impedance state. A delayed signal from the $\overline{RAS}$ generated clock and the $\overline{RAS}$ signal itself is gated with $\overline{CAS}$ to ensure that the $\overline{CAS}$ generated clocks do not commence until the optimum time and while $\overline{RAS}$ is active. The $\overline{CAS}$ generated clock latches the column address which selects the appropriate column of the memory array. The data from the selected sense amplifier is transferred to the output buffer within an access time (t_{CAC}) after $\overline{CAS}$ period (t_{RP}). The total time taken by the active and inactive period of $\overline{RAS}$ represents the memory cycle which is often referred to as the random read cycle time.

3.2 Normal write cycle

During a normal read cycle, the write enable (WE) signal is inactive, but for write operation it is activated at some time (t_{WCS})prior to the $\overline{CAS}$ active edge — see Fig. 16(b). Until this point the access cycle is exactly as the read cycle already described. Once the $\overline{WE}$ signal is activated the data set up at the data input pin (t_{DS} applies) is latched into the chip on the $\overline{CAS}$ active edge. The data is held valid for a

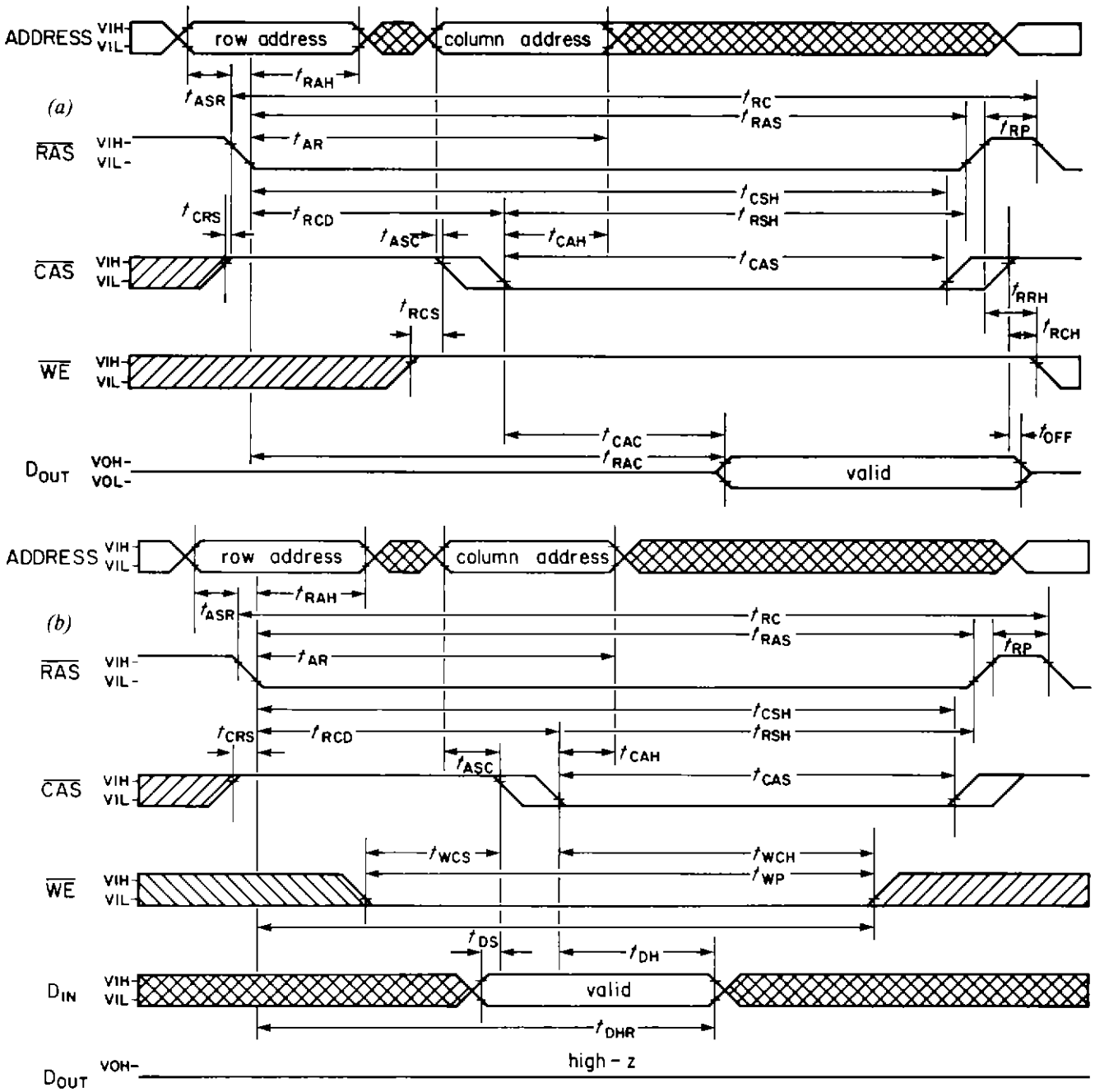


Fig. 16 - Dynamic read-write memory signal waveforms: (a) read cycle (b) write cycle

period (t_{DH}) after this. The data is written into the selected sense amplifier and the selected cell. During a normal write cycle the data output buffer is disabled.

This latter property together with the absence of a data output latch makes it possible to connect the data input and output pins together provided that only normal read and write cycles are used. This is referred to again in Section 3.7.

3.3 Other forms of write cycle

Another form of write cycle is common to all the most recent generations of dynamic read-write memory devices. It is the 'read-modify-write' cycle whereby an addressed cell can be accessed to read its data and after which different data is written to the same address. Typical waveforms associated with the read-modify-write cycle are shown in Fig. 17. When the $\overline{WE}$ signal is delayed beyond the falling edge of $\overline{CAS}$ by a prescribed minimum period (t_{CWD}) the data output will contain data read from the selected cell in the same way as for a read cycle. Data to be written into this cell is set up and held relative to the falling $\overline{WE}$ edge which now directly performs the write function.

If it is not necessary to make use of the read data during this cycle, another alternative to the normal write cycle is one where writing can occur

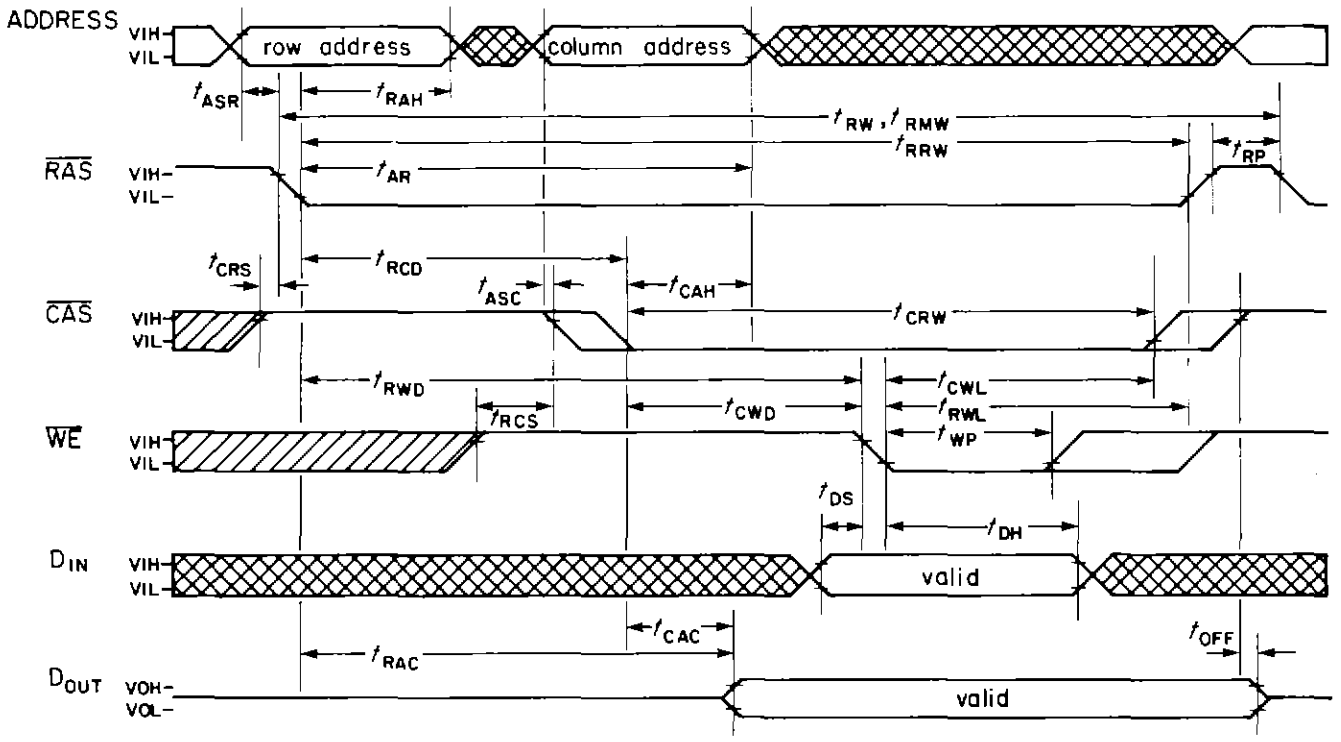


Fig. 17 - Dynamic read-write memory late-write/read-modify-write cycle

'late'. A cell is addressed in the same manner as a read-modify-write cycle and the read data appears on the data output pin. However, the $\overline{WE}$ signal may now be applied before the data output is valid to shorten the cycle length (t_{RW}). Input data is still applied referenced to the $\overline{WE}$ falling edge and the waveforms of Fig. 17 are still valid. It is sometimes more convenient, in a particular application, to use the 'late write' form rather than 'early' (normal) write. This might be done, for instance, in high-speed shift registers when data for writing does not become valid until after the memory address has been applied.

3.4 Attempts to speed up the cycle

The penalty of slower access time caused by multiplexed addressing has led manufacturers to find ways of increasing the speed of operation in some circumstances. One means of increasing speed without increasing operating power is possible provided that successive memory operations occur at locations sharing the same row address. This is known as 'page mode' and a typical read and write cycle is shown in Fig 18. The row and column components of an address are applied in the normal way and, depending on the polarity of $\overline{WE}$, data is either written to or read from the selected cell. However, if $\overline{RAS}$ is maintained active, when $\overline{CAS}$ is made inactive, the data for the whole of the addressed row remains available on the sense amplifiers for that row. By applying a second column address and a second $\overline{CAS}$, in the case of read operation, another sense amplifier can be selected and its data transferred to the output buffer without having to re-address the row again. The whole of the cells in a row may be accessed in the same way and similarly for a repeated write operation. Successive accesses can therefore be repeated at very much shorter intervals than is the case for normal read or write cycles. Typically a speed increase of the order of 30% is possible. Besides normal read and write cycles, read-modify-write cycles can also be used in page mode.

An improvement over page mode operation for random access requirements is achieved in some devices by accessing more than one cell at a time from a single applied address. In the form developed, four cells at successive column addresses are accessed in what has become known as 'nibble' mode. (The term 'nibble' is borrowed from computer terminology where it generally means half-a-byte, i.e. half of eight bits). The resulting waveforms for a 64 K DRAM (on which the nibble mode first became available) are given in Fig. 19 which describes the read and write cycles. The first row and column address supplied, determines the address of the first cell in the sequence of four. Toggling $\overline{CAS}$ causes the next three successive cells to be accessed. If a fourth active $\overline{CAS}$ is applied in the same sequence, the sequence of selected cells repeats. The nibble mode read and write minimum cycle time (t_{NC}) for currently available 64 K DRAMs is 55 ns.

Two new features have been introduced on the latest 256 K CMOS devices. Transparent, i.e. level-

triggered rather than edge-triggered, row address latches allow a much shorter row address capture window which reduces the row address hold time (t_{RAH}). Also the column address decoding is now static so that no address strobe is required to select an individual cell in any row. Once a row has been selected, the column addresses may be freely changed and the output data follows them.

3.5 Power consumption

The dynamic circuitry causes most operating current to be drawn on address strobe edges. Thus, the operating power is primarily a function of operating frequency, i.e. the speed at which consecutive memory cycles occur. To a secondary extent the operating power also depends on the electrical loading of the data output connection. Typical current waveforms for a 16 K DRAM (which employs three voltage levels) operating with different cycles are shown in Fig. 20. It has been estimated that about 60% of the operational power is due to $\overline{RAS}$ and the remainder to $\overline{CAS}$. The reduction of operating current with reduced operating frequency is shown in Fig. 21. At its minimum operating random read-write cycle length of 375 ns the maximum current drawn at 12 V for the mid-speed option (suffix-3) measures 35 mA but increasing the cycle length to 1 μs, it is reduced to

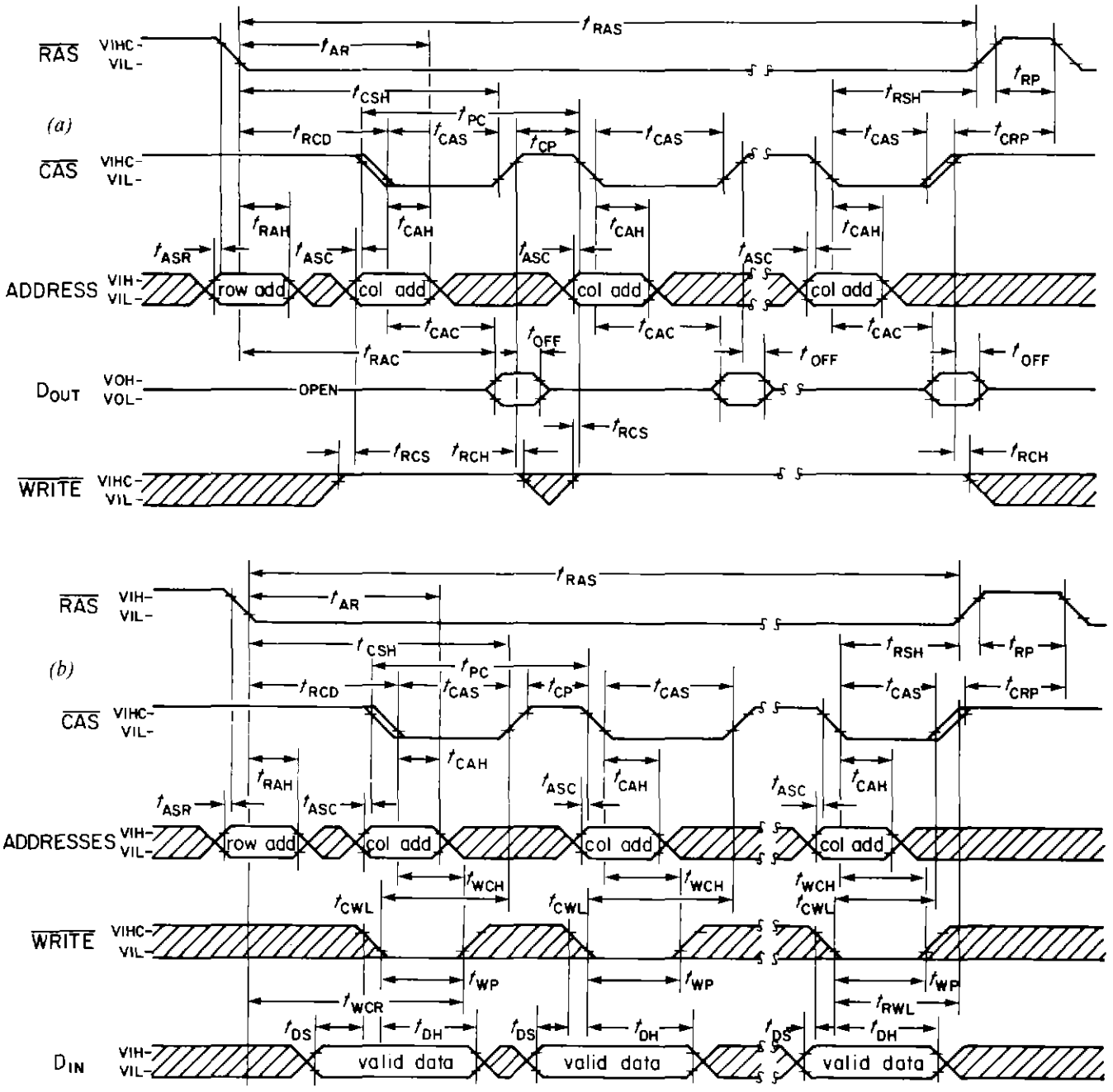


Fig. 18 - Dynamic read-write memory signal waveforms: (a) page mode read cycle (b) page mode write cycle

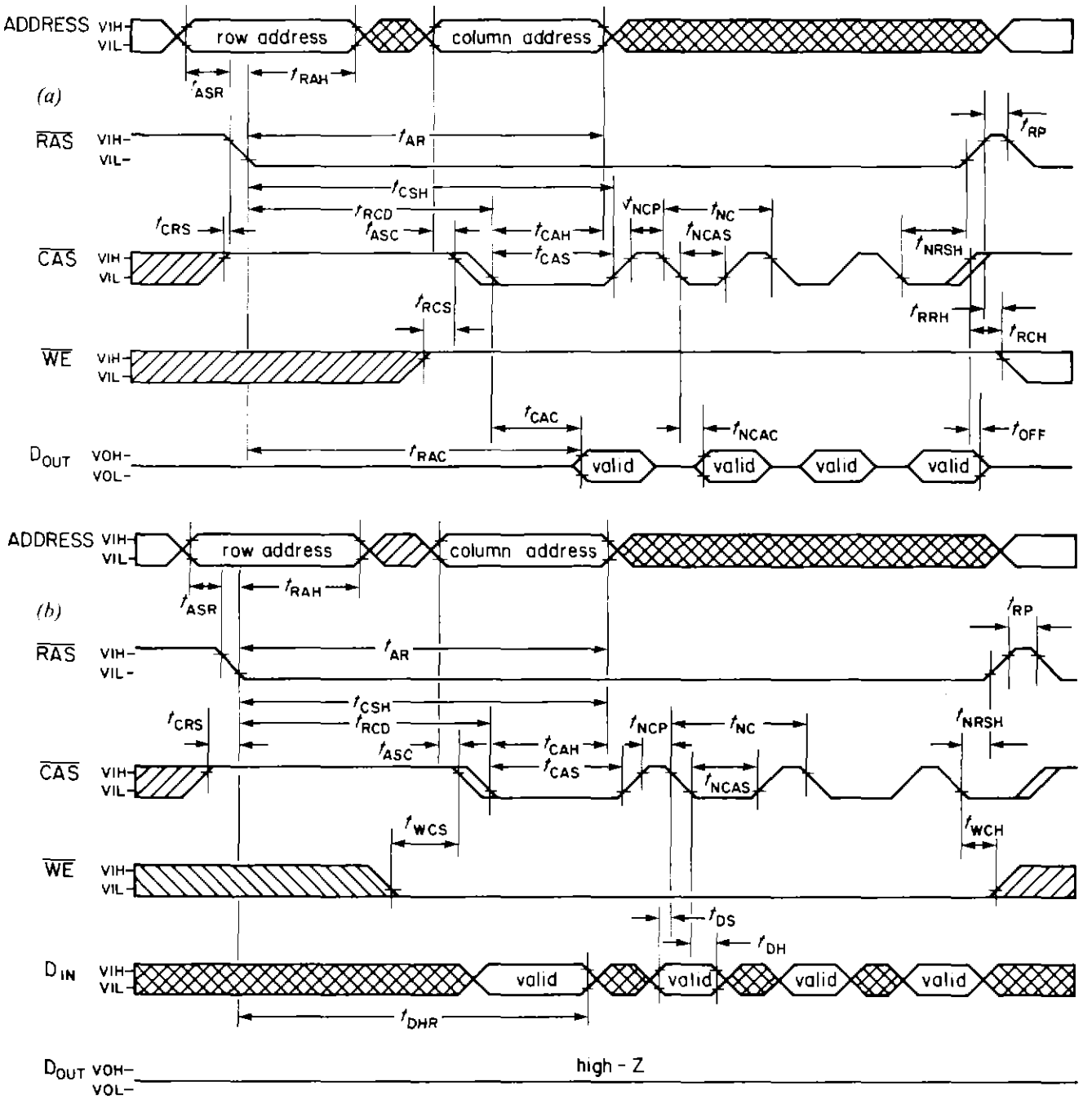


Fig. 19 - Dynamic read-write memory signal waveforms: (a) nibble mode read cycle (b) nibble mode write cycle

20 mA — a significant saving! The minimum overall system power consumption is achieved if $\overline{RAS}$ is used to chip select devices because unselected chips then revert to operation in the low-power standby mode regardless of $\overline{CAS}$.

3.6 Power distribution

The transient operating currents shown in Fig. 20 can cause significant power rail and ground noise unless precautions are taken with the distribution of power to individual devices and adequate decoupling is provided. The power conductors and ground conductors should ideally be fully 'gridded' to minimise their impedance and reduce the amplitude of noise on these lines which can otherwise erode signal margins. The term 'gridding' means using power conductors interconnected orthogonally in the form of a lattice. Adequate decoupling may be provided by a 0.1 μF ceramic capacitor, connected as directly as possible between the power and ground pins of each device, to suppress high-frequency transients. Also, a larger tantalum capacitor, say 47 μF, should be placed near the edge connector of the memory board where the power lines connect to the motherboard. This provides the bulk energy storage required to prevent an unacceptable voltage drop due to the main power supply being remote from the memory board at the

end of a relatively long inductive path. In the earlier triple voltage level devices it was also important for the substrate bias supply to be applied first and removed last; otherwise it was possible to cause catastrophic failure of some parts of the device by semiconductor junctions becoming effectively short-circuited.

3.7 Data output control

It has been common practice for the data output buffer of dynamic read-write memory devices to be controlled by signals derived from the applied $\overline{CAS}$. The block diagram of Fig. 15 illustrates the hardware implementation of this which has also applied to devices from the 16 K generation onwards. In these cases, whenever $\overline{CAS}$ is high the data output is unconditionally high impedance — see Fig. 16. When $\overline{CAS}$ is activated and $\overline{WR}$ is held high the data output pin becomes active after the appropriate access period and contains the data read from the selected cell. This applies equally to the read, late-write and read-modify-write cycles. In a normal write cycle the data output remains high impedance because the active $\overline{WR}$ signal which preceeds the active $\overline{CAS}$ edge overrides the enabling action of $\overline{CAS}$. If the device operation is restricted to normal read and write modes it is possible to connect the separate data input and output pins together to form a common data input-output bus.

The data appearing on the data output pin during a read cycle is not normally latched within the device but remains valid until the end of the active $\overline{CAS}$ pulse. During this time there is the opportunity to latch the data using external circuitry. This method of operation, which has again been common practice,

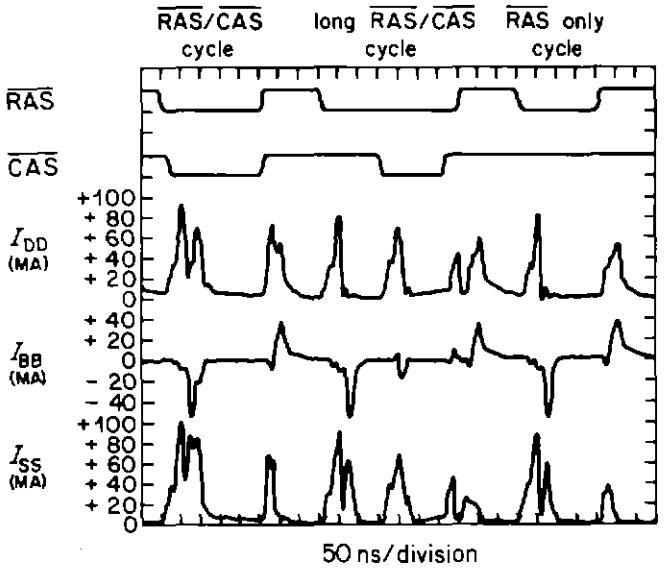


Fig. 20 - Dynamic read-write memory: typical current waveforms.

allows devices from the 16 K generation onwards to have their output pins interconnected to form a common output data bus without the need for the data from unselected devices, sharing the data bus, to be turned off.

A more recent innovation has appeared in the 16 K x 4 arrangement of 64 K DRAMs offered by a number of manufacturers. An output enable, $\overline{OE}$, function provides an extra level of output control which allows a common input-output bus even in the read-modify-write mode and in this device the data input and output pins are connected internally.

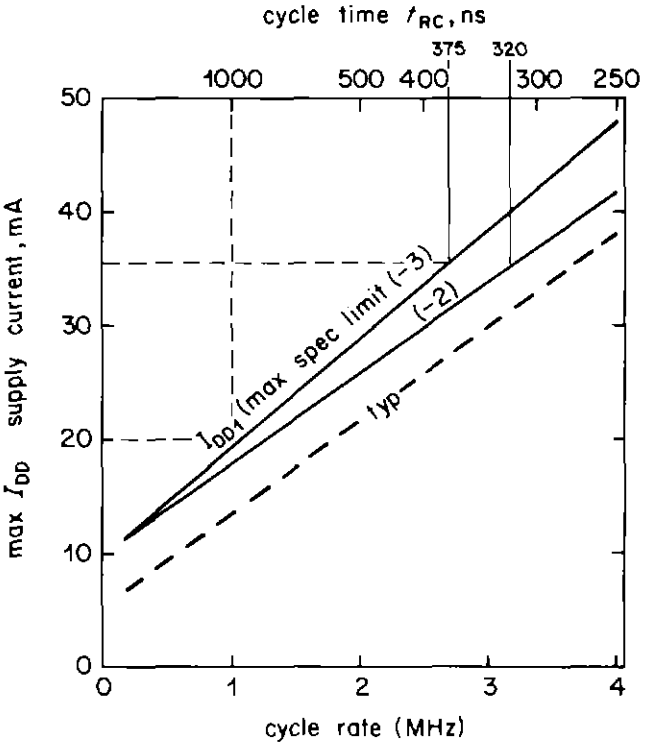


Fig. 21 - Dynamic read-write memory: current consumption characteristic.

3.8 Refresh

The volatile nature of dynamic read-write memory devices makes it necessary to refresh the data stored within the capacitative cells before the charge decays to such a degree that data is lost. Each succeeding generation of DRAM has tried to maintain some level of compatibility with its predecessor as far as the refresh requirements are concerned. Thus, for example, when 64 K devices were introduced, some manufacturers provided a 128-cycle refresh every 2 ms to match the 16 K devices and others, seeking to halve the number of sense amplifiers within the chip, provided a 256-cycle refresh every 4 ms. No manufacturer provided both options on one chip because that would have increased the die size to include the extra sensing amplifiers and option selection mechanism.

Apart from any special provision that may be made to meet refresh requirements, any type of memory cycle which accesses a row of the memory matrix causes all the cells within that row to be refreshed. There are other ways in which refresh can be achieved. It is sufficient to operate a $\overline{RAS}$ only cycle to perform a refresh operation and because no $\overline{CAS}$ is required there is a significant power saving. It is used in all generations from 4 K through to the latest 256 K DRAMs and memory addresses are supplied externally. A $\overline{CAS}$-before-$\overline{RAS}$ method, used by some manufacturers for the 64 K and 256 K devices, avoids the need to provide external memory addresses, with resulting savings of power and board space. $\overline{CAS}$ is brought low before $\overline{RAS}$ and this trigger advances an internal counter which provides the refresh row address: the address pins of the device are ignored. Only refresh is available in this mode and no data can be written or read. No device selection occurs and the data output pins of each device remain unchanged. This means that if this type of refresh is applied directly after a read cycle, for instance, the data output is maintained and the refresh action is effectively 'hidden'. On some 64 K DRAMs this internal refresh function can be initiated by applying a signal to one of the pins of the dual-in-line package which is otherwise unallocated, No separate $\overline{CAS}$ is therefore required.

3.9 Interfacing

As the operating frequency of memory devices increases it becomes important to reduce the propagation delay and 'ringing' of applied addresses and the other signals because of the capacitative loading which the devices present. In practice the situation is far from that of an ideal transmission line. Typically a driving buffer supplying signals to a string of memory devices along a printed-circuit board (p.c.b) track might introduce large overshoots by the time the signal reaches the last device. The signal waveform can be improved by either including a small series resistor between the driver and first device in an attempt to match the source impedance of the driver to the printed-circuit board track impedance or by terminating the transmission line directly in a simple RC network.

3.10 Reliability

The results of tests carried out by Hitachi and published in their current semiconductor memory data book indicate that MOS memories are very reliable devices. At elevated temperatures (up to 150 °C ambient) the failure rate was measured as less than 1 in 10^5 out of a batch of devices tested over a period of over a million component hours. Tests at 85% relative humidity indicate a similar failure rate. No failures were detected due to thermal cycling (between -55 °C and +150 °C), soldering heat (260 °C for 10 seconds), mechanical shock (1500 g for 0.5 ms), variable frequency (20 Hz to 2 kHz) or constant acceleration (20,000 g).

3.11 Comparison with static memory devices

The faster access time of static memory devices owes much to the direct addressing of the memory cell matrix. Read cycles operate in a completely static mode in that no external clocks are required to access the stored data. This is accomplished by a sense address transition circuit which initiates an internal clock, wherever a change occurs in the logical state of the address lines. The static loads associated with the static sense amplifier circuitry account for a steady current drain in these devices. A comparison of current drawn in dynamic and static devices is presented in Fig. 22. For static devices the current waveform is more dependent on the active duty cycle and a greater overall power is dissipated.

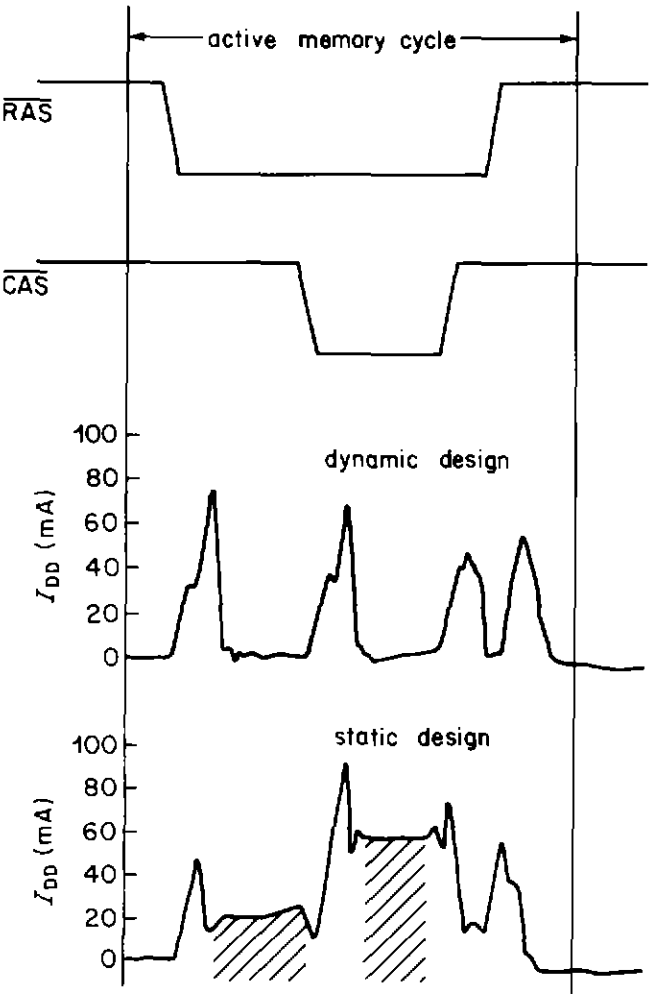


Fig. 22 - A comparison of the current consumption characteristic in static and dynamic read-write memory devices

4. DESIGN PHILOSOPHY

4.1 Introduction

The main choice, for the majority of applications to television engineering, is that between the use of static or dynamic semiconductor memory. Compared with other technologies, dynamic memories represent a very attractive and cost-effective solution for large random-access, mass storage applications where speed is not so important and where cost and overall power consumption considerations dominate.

Static memory devices are suitable for smaller storage units (say up to about 64 Kbits) where the speed advantage is beneficial and the higher power consumption and cost can be tolerated. Thus, broadly speaking, dynamic memories tend to be used almost exclusively for building stores of television picture and multi-picture capacity whereas static devices are more appropriate for television line stores and microprocessor memory. For very small stores, such as that required for delaying a video signal by a few sample periods for example, the even faster bipolar devices are useful.

The use of semiconductor memory generally falls into one of two categories, namely, that which serves as a delay and that which offers random access of a storage block. The first requires a relatively simple means of control in which the memory address is continually incremented for a period defining the delay and then reset to the start address. For random access a means must be provided to generate the store address and supply the necessary write enable (WE) polarity for either read or write cycles.

The main questions to answer when designing a large dynamic semiconductor-memory-based random access store are:-

a) What total store size is required (in terms of storage capacity and the number of bits to define each data sample)?

b) Which types of memory chips are available? Are there special features e.g. page mode, nibble mode?

c) What multiplexing arrangements are required to accommodate the fastest data transfer rate?

d) How many independent read or write access ports are required?

e) How many memory chips and their support chips can be satisfactorily housed on a single printed circuit board?

Another question concerns the refresh requirements of dynamic memories. In television picture storage applications, the memory chips can be arranged to be accessed sufficiently frequently during normal video read cycles to service the refresh function and generally no special arrangements are necessary.

Some of the important design parameters raised by the questions above are considered in more detail in the remainder of this Section.

4.2 Store size

The number of data samples required to support one television picture depends on the digital video sampling frequency and the television standard used. Fig. 23 shows the number of data samples applicable to two television standards, System I (UK) and System M (USA), for a range of sampling frequencies between 12 MHz and 20 MHz. It is often sufficient to store the active picture area only and to omit those samples which occur during the field- and line-blanking intervals and both cases are presented in the figure. It is worth noting that at the sampling

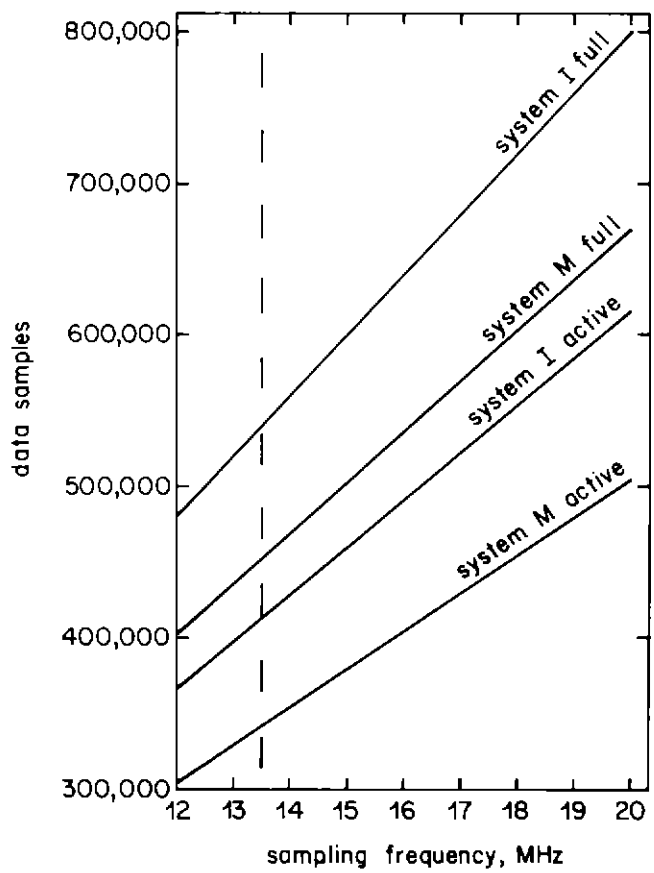


Fig. 23 - The number of data samples generated, in television pictures based on the System I and System M television standards, as a function of digital sampling frequency.

frequency of 13.5 MHz, shown dotted, and which was to become an International Digital Television standard[9], more than 512 K samples are required to hold a single picture including blanking periods. This is slightly inconvenient in terms of the number of memory devices required to store this many samples because of the 'powers of two' factor governing memory device package sizes.

4.3 Multiplex factor

The relatively slow speed of dynamic memory devices can be matched to video data rates by demultiplexing the data by a factor depending on the ratio of these quantities. There are basically two ways of doing this. Firstly, the data may be sequentially distributed like the action of a commutator as shown in Fig. 24(a) (DMX) and retrieved from the memory devices in similar fashion (MX). An advantage of this method is that a minimum delay can be achieved, but a disadvantage is that multi-phase clocks and addresses are required. Alternatively the incoming data can be assembled into blocks and presented to the stores simultaneously as shown in Fig. 24(b) and on reading, the retrieved blocks dispersed. An advantage of this method is that only one clock and address phase is required. With either method the amount of delay is quantised into units of F clock periods where F is the demultiplex factor. Finer delay trimming can be obtained using a small buffer store.

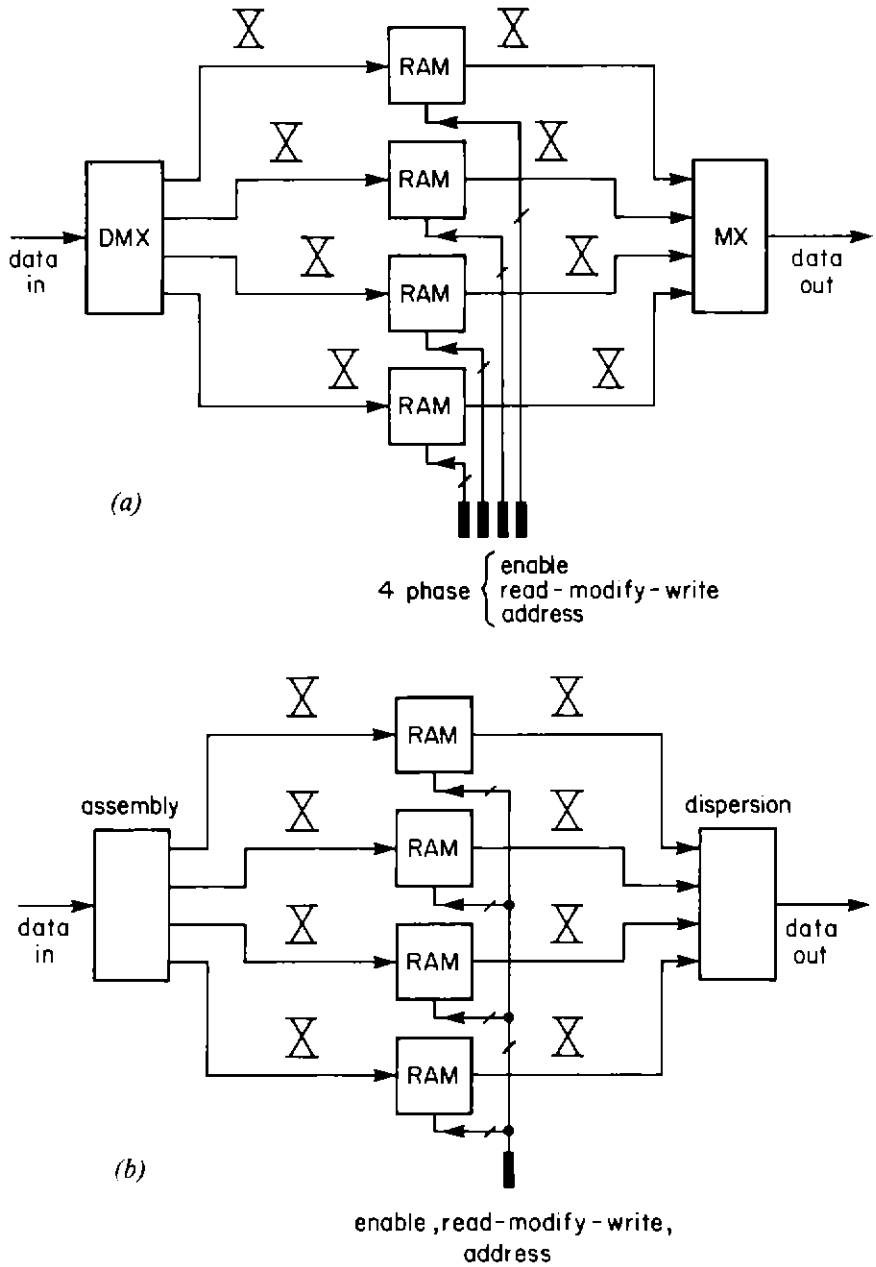


Fig. 24 - Two methods of multiplexing semiconductor memory devices: (a) sequential distribution (b) simultaneous distribution

4.4 Store configuration

A number of television picture stores have been constructed based on the schematic shown in Fig. 25(a) for one bit of data. This arrangement splits the storage into two separate sections, labelled A and B, which are operated on an alternate write and read cycle — see Fig. 25(b) — (dynamic memory devices cannot accept simultaneous write and read addresses). The input data is demultiplexed by a convenient factor F for which the minimum value depends on the data sampling rate and the minimum memory cycle time. The maximum value for F depends on the overall size of the store which, in turn, determines the maximum package count. In general there are n sub-sections in each half of the store and each sub-section contains F packages. In the example shown, $n = 4$ and the store can be considered as four separate stores 'in parallel' sharing a common data input port but each feeding a separate data output port. In this case, four separate output ports can be provided, one from each pair of sub-sections taken from A and B as shown. For standards conversion[10], for example, four separate outputs containing data on four successive television lines can be provided by such a store arrangement. Moreover, the store control is relatively straightforward requiring an input F-way serial-parallel converter, an F-way parallel-serial converter and simple address generators driving each section independently.

The maximum number of access ports for this arrangement is $2n$ but because of the alternate write-read operation, there can be only n independent ports used for data output. In this configuration the store cycle length is defined to be $2F$ clock periods and so the store delay is quantised into units of the same amount.

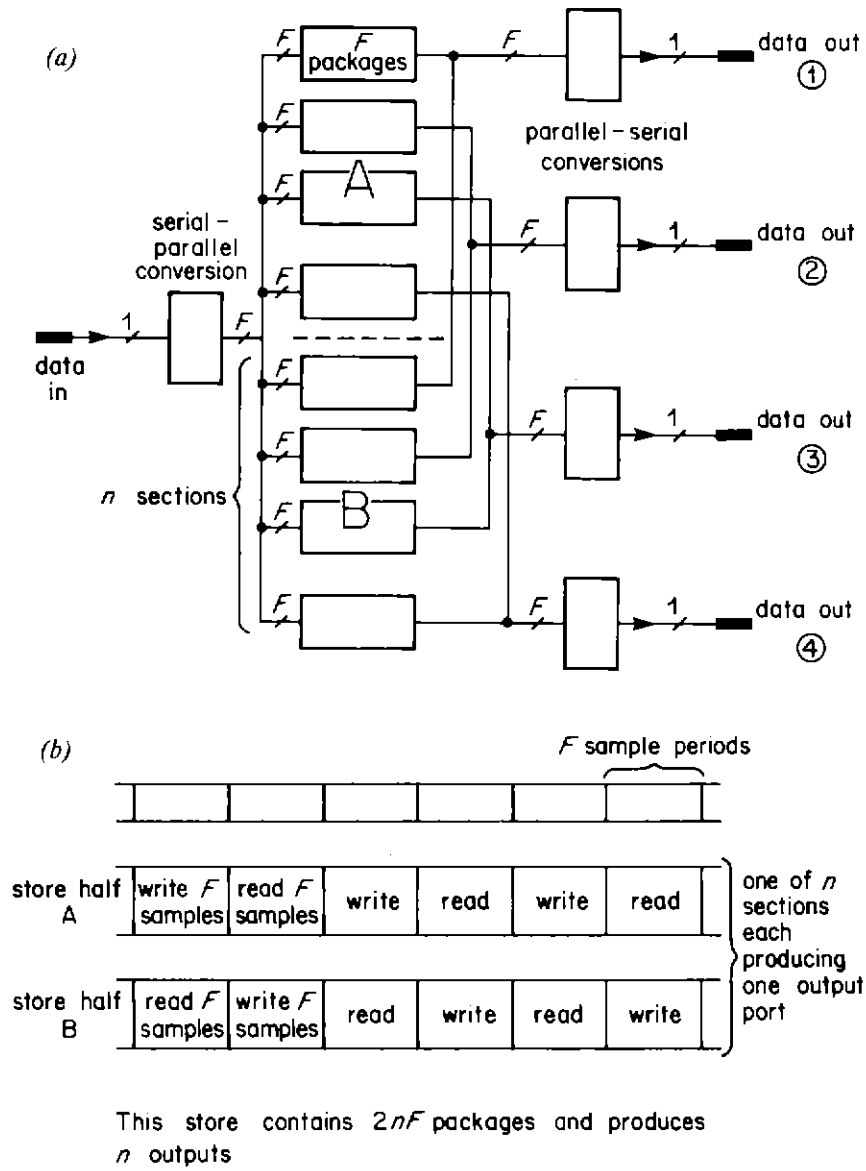
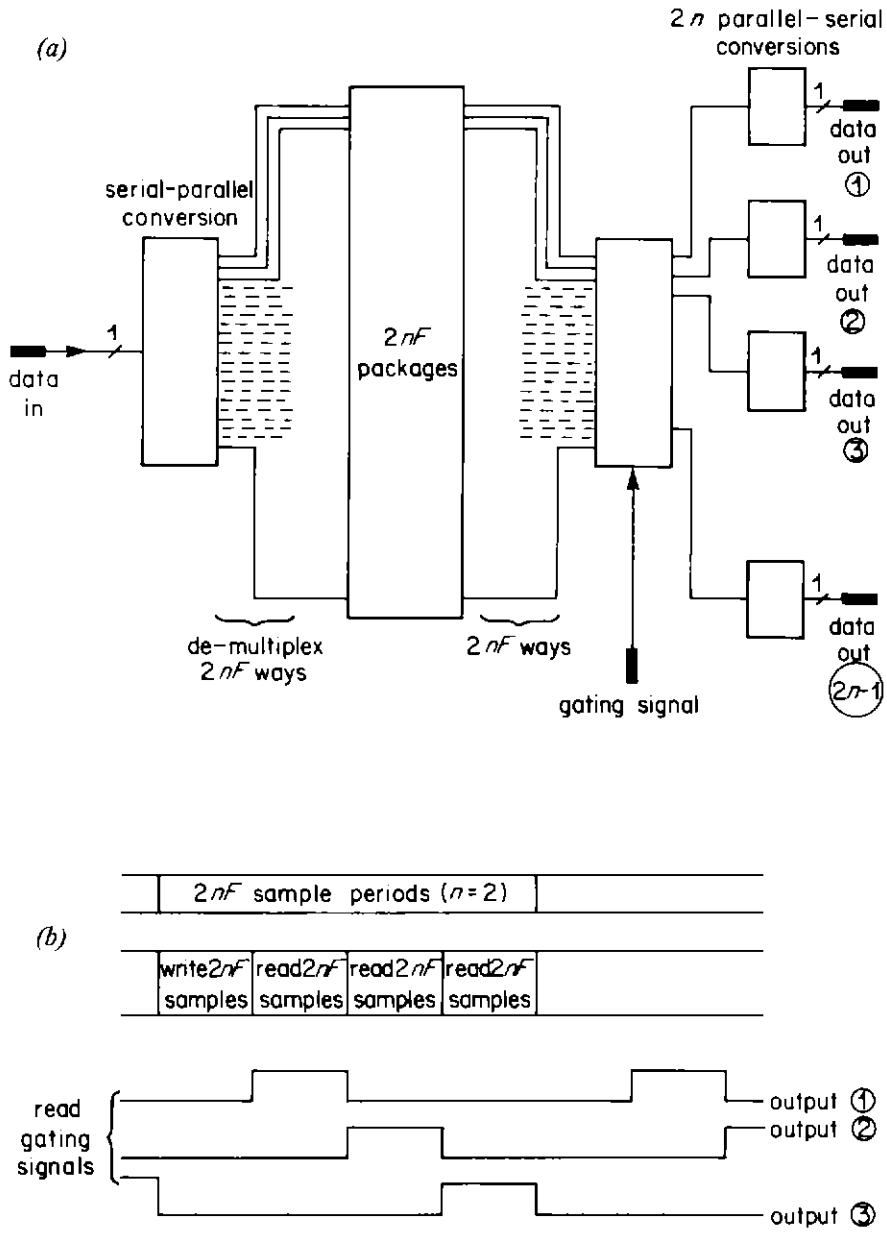


Fig. 25 - Conventional 'split' store with A and B sections for alternate reading and writing: (a) block schematic (b) alternate read/write sequence

An alternative store arrangement is shown in Fig. 26(a). The demultiplexing factor is now much greater than in the previous case and can take any value between F and $2nF$; the figure is drawn for the maximum value of $2nF$. The store cycle is now $2nF$ sample periods long and a timing diagram (Fig. 26(b)) is given for a store with $n = 2$. Because F sample periods are required for each device read or write cycle it is possible to access the store four times independently within the store cycle. The first is used, in this example, to write a block of $4F$ samples, one into each package. In the second to fourth cycles $4F$ samples of data are read from a different address each cycle. The output data is off-loaded onto a common data bus and gating signals must be applied to separate the data destined for separate output ports. The control is therefore more complex than in the previous store arrangement, requiring larger serial-parallel and parallel-serial converters and also the generation of output gating signals to separate the output data. The advantage of this arrangement,however, is that three outputs can be provided or, in general, a maximum of $2n$ (with no writing cycle), which is double that of the previous arrangement.



This store contains $2nF$ packages. There can be $2n$ independent access ports, all of which may be outputs. The example here shows one input and 3 outputs for $n = 2$.

Fig. 26 - An alternative store arrangement with gated outputs: (a) block schematic (b) read/write sequence and gating waveforms

5. CONCLUSIONS

An historical introduction has been presented in order to explain current trends in semiconductor technology development. Dynamic memory devices continue to evolve with each succeeding generation, quadrupling the memory capacity within a relatively steady die size. The increased memory cell density causes cell sizes to shrink and now minimum line widths of less than 1 μm are in prospect. Power consumption for each cell has reached the 1 μW/bit level and the cost of each cell is less than one-thousandth of a penny at today's prices. Static devices have generally followed their dynamic counterparts at each stage of technological advance and offer a speed advantage at the expense of greater power consumption and cost.

Semiconductor memory devices have steadily become easier to use with the need for a single power supply only and relaxed operating margins. Built-in refresh mechanisms have reduced the disadvantage inherent in dynamic devices. Improved data input and output control has resulted in a greater range of operating modes.

Attention has been drawn to the main questions to be answered when designing random access stores based on semiconductor memory devices. These include the definition of the total store capacity and the multiplexing arrangements, in order to match the required data transfer rate and to accommodate multiple store access.

6. REFERENCES

1. RILEY, J.L. 1987. A Review of the Semiconductor Storage of Television Signals: Part 2 — Applications 1975 - 1986. BBC Research Department Report No. BBC RD 1987/6.

2. AHLQUIST, et al. 1976. A 16384-bit Dynamic RAM. IEEE Jnl. SSC, Vol. SC-11, No. 5, Oct 1976.

3. MAY, T.C. and Woods, M.H. 1979. Alpha-particle-induced Soft Errors in Dynamic Memories. IEEE Trans. on Electron Devices, Vol. ED-26, No. 1, pp 2-9, Jan. 1979.

4. CHAN, J.Y. et al. 1980. A 100 ns 5 V only 64K x 1 MOS Dynamic RAM. IEEE Jnl. SSC, Vol. SC-15, No. 5 pp 839-46, Oct 1980.

5. SHIMOHIGASHI, K et al. 1982. An N-well CMOS Dynamic RAM. IEEE Trans. on Electron Devices, Vol. ED-29, No. 4 pp 714-18, April 1982.

6. KUMAGAI, et al. 1983. A 64 K CMOS RAM with Low-power Circuit Technology. IEEE ISSCC Digest of Technical Papers, Feb 1983.

7. CHWANG, R et al. 1983. A 70 ns High-quality CMOS DRAM. IEEE ISSCC Digest of Technical Papers, pp 56-57, Feb. 1983.

8. BABA, F et al. 1983. A 35 ns 64K static column DRAM. IEEE ISSCC Digest of Technical Papers, pp 64-65, Feb. 1983.

9. Encoding Parameters of Digital Television for Studios. CCIR Recommendation 601, XVth Plenary Assembly, Geneva 1982. Vol.XI, Part 1, pp 271-273.

10. ROE, G.D. and DALTON, C.J. 1979. Method of an Apparatus for Processing Television Signals. UK Patent Specification GB 2013067.

Printed by BBC RESEARCH DEPARTMENT, Kingswood Warren, Tadworth, Surrey, KT20 6NP